DISPATCH

Genome Evolution: Mutation is the Main Driver of Genome Size in Prokaryotes

Gabriel A.B. Marais[1], Bérénice Batut[2], and Vincent Daubin[1]

[1]Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Évolutive UMR 5558, F-69622 Villeurbanne, France

[2]Albert-Ludwigs-University Freiburg, Department of Computer Science, 79110 Freiburg, Germany

Summary

Despite intense research on genome architecture since the 2000's, genome-size evolution in prokaryotes has remained puzzling. Using a phylogenetic approach, a new study found that increased mutation rate is associated with gene loss and reduced genome size in prokaryotes.

In 2003 [1] and later in 2007 in his book "The Origins of Genome Architecture" [2], Lynch developed his influential theory that a genome's complexity, represented by its size, is primarily the result of genetic drift. More genetic drift would result in larger genomes. This view relies on population genetic models as well as ***(AU:OK?)*** an analysis comparing prokaryotes and eukaryotes. On one side, prokaryotes undergo little genetic drift as they typically have large effective population sizes (*Ne*) and small genomes. On the other side of

the spectrum, multicellular eukaryotes have large genomes. Because of their small *Ne* and a large amount of drift, selection is inefficient in purging the genome of genomic parasites, such as transposable elements, and slightly deleterious indel mutations such as gene duplications. As a consequence, their genome sizes increase. Unicellular eukaryotes are found in between. Immediately after these publications, it was pointed out that the theory of 'mutational hazard,' as it is now referred to, does not sufficiently explain ***(AU:OK?)*** the diversity of genome size within Bacteria and Archaea (for example, see [3]). In particular, bacterial endosymbionts such as *Buchnera aphidicola* living in aphids have very small *Ne* (similar to that of their obligatory hosts) and are among the smallest known bacterial genomes. In 2009, Kuo *et al.* [4] showed that the relationship between genome size and *Ne* is the reverse in bacteria: that is, bacteria with small *Ne* have small genomes. Bacterial genomes, even large ones, usually contain very few transposable elements, and genome size is mainly driven by gene content. In small *Ne* bacteria such as *Buchnera*, the selective pressure to keep genes is low. Many non-essential genes become non-functional and are deleted, resulting in organisms with few genes and thus small genomes. However, Kuo *et al.*'s study did not include the mysterious marine bacteria *Prochlorococcus*. In the early 2000's, marine biologists discovered that *Prochlorococcus*, one the most abundant free-living bacteria on earth, has a reduced genome compared to its relative, *Synechococcus* (for example [5,6]). Given its abundance, genome reduction in *Prochlorococcus* cannot be explained by a small *Ne*. A number of hypotheses have been proposed to fit *Prochlorococcus* into the general picture [7], including an adaptation to a marine environment deficient in nitrogen and phosphorus (the genome streamlining hypothesis) or an increase in mutation rates. Interestingly, *Prochlorococcus* genomes lack key DNA-repair genes, and increased mutation rates can result in gene loss ***(AU:OK?)*** under weak selection [8]. These hypotheses, however, have remained untested, and our understanding of the evolution of genome size in bacteria still has to deal with a

number of inconsistencies [7]. In this issue of *Current Biology*, Bourguignon *et al.* [9] provide for the first time strong evidence that the main driver of genome size in bacteria is mutation rate, both in free-living bacteria and in bacterial endosymbionts [9]. This study is a first step towards a new theory for explaining genome size in prokaryotes.

To disentangle the effects of *Ne* and mutation rates, Bourguignon *et al.* have updated a classical approach using the rate of nonsynonymous substitutions (*dN*) and the rate of synonymous substitutions (*dS*) in a set of conserved genes. *dN* is affected by mutation rates and the intensity of selection (which depends on *Ne*). In contrast, *dS* is affected mainly by mutation rates (neglecting selection on synonymous codon usage. Looking at *dN* and *dS* together provides information about***(AU:OK?)*** mutation rates, whereas looking at the classical *dN/dS* ratio provides insights into selection and *Ne*. This approach has been used in the past (for example by [4]) but very distant species were compared, which can be misleading. Here, Bourguignon *et al.* [9] have sequenced new genomes and collected some from public databases to assemble a very large dataset comprising nine bacterial and archaeal lineages, with seven being free-living, including the marine cyanobacteria *Prochlorococcus* and *Synechococcus.* The authors analyzed each lineage separately and compared close relatives with varying genome size, an analysis never done before. In this paper, they restricted their analysis to sets of genes shared by all species of a lineage undergoing supposedly similar selection, thus making *dN/dS* a good proxy for *Ne*. In addition, using a model for gene content evolution, they could infer gene loss and gain for each branch of the phylogenetic trees. For the marine cyanobacteria, they used the alignment of 31 core genes to estimate *dS* and *dN* along the branches. Dividing the estimates by branch length, they obtained gene loss, *dS*, and ***(AU: is this what you mean? Each parameter measured per unit time?)*** *dN* per unit of time and found out that gene loss is correlated with *dS* (strongly) and *dN* (moderately), but is not correlated with *dN/dS*. Their result obtained on a small number of

genes is however robust to a number of potential biases: phylogenetic inertia ***(AU:OK?)***(as this can lead to spurious correlations), G+C content (which can affect *dS* and *dN* estimates), codon bias (as selection on codon bias decreases *dS*), and moreover, they used residuals from linear regression instead of ratio (as this can affect the output of statistical tests). Genome reduction in *Prochlorococcus* is then predominantly driven by an increased mutation rate.

But is this effect specific to a lineage of photosynthetic bacteria that has to cope with a deprived and mutagenic environment? Apparently not. In the other six free-living lineages, similar results are found in four of them, *Thermococcus, Corynobacterium, Micrococcineae, Flavobactericeae,* which represent very contrasting lifestyles and ecology. In the remaining two lineages (*Gammaproteobacteria, Mycobacteria+Nocardiaceae*), the results were different. They either found no correlation at all (*Gammaproteobacteria*) or a correlation between gene loss and *dN/dS* (*Mycobacteria+Nocardiaceae*), suggesting that in the latter ***(AU:OK?)*** lineage *Ne* does matter with respect to genome size. Hence, the association between mutation rates and genome size clearly applies to free-living prokaryotes, albeit not universally. And what about endosymbionts? Could mutation rate also play a role in genome reduction? The study included two endosymbiont lineages: *Blattabacterium* and the iconic *Buchnera.* To estimate the gene content evolution they used a different model with gene loss only—as gene gain by horizontal transfer is extremely rare in intracellular endosymbiont bacteria—and inferred gene loss along the branch of the trees. ***(AU: preceding sentence restructuring ok? Did I understand your intent correctly?)*** From a set of 353 core genes, they also estimated *dS* and *dN* using the same methods as explained before and obtained per unit of time estimates of gene loss, *dS* and *dN*. They found that gene loss is strongly correlated with *dS*, moderately with *dN*, and very weakly with *dN/dS*. In most of the controls they used ***(AU:OK?)*** the correlation between gene loss and mutation rate (*dS* and *dN*) remained, whereas the correlation between gene loss and *Ne* (*dN/dS*) disappeared.

These findings suggest that mutation rate plays a central role in the evolution of genome size in prokaryotes, both in free-living bacteria and endosymbionts. This represents ***(AU:OK?)*** a substantial change in our view on how genome size evolves in prokaryotes in general and in endosymbionts in particular. It is well known that mutation rates are determined by the presence of DNA-repair genes. Now, the data on DNA-repair genes that Bourguignon *et al.* obtained (along with others from the literature; for example, [10]) ***(AU:OK?)*** clearly indicate that genome size and the number of DNA repair genes are correlated. As DNA-repair genes are lost, mutation increases and counteracts relatively weak selection for gene function ***(AU:OK?)***, resulting in gene loss and genome reduction in prokaryotes [8]. Why DNA-repair genes are lost in the first place remains to be explained and may correspond to different situations. For instance, a large body of work has shown that mutator strains (with increased mutation rates) can adapt faster than wild-type strains in a changing environment, simply because advantageous mutations arise faster in the former [7,8]. In this scenario, mutation rate is the main target of selection and genome reduction is just a side effect. In general though, the initial mutation rate can be restored by recombination or gene transfer once an adaptive solution to the environment has been found. In endosymbionts however, the genetic material for restoring a functional repair machinery may not be available because the symbiont does not come into contact with other bacteria, hence initiating an irreversible decay of the genome. It is also possible that a streamlined genome is beneficial in some environments, such as nitrogen- and phosphorus-poor tropical waters [5,6]. In this case, genome size could be the main target, and the loss of DNA-repair genes is just a means to an end. Losing DNA-repair genes could also be due to drift, especially at the initial stages of endosymbiosis, but the general lack of correlation between gene loss and *Ne* does not support this as a general explanation. More work is needed to test the mutator versus genome streamlining hypotheses and the consequences for genome evolution in Bacteria and Archaea ***(AU:OK?)***. However,

this study is certainly a major step towards the development of a comprehensive theory of prokaryote genome-size evolution.

References

1. Lynch, M., and Conery, J.S. (2003). The origins of genome complexity. Science *302*, 1401–1404.

2. Lynch, M. (2007). The origins of genome architecture (Sunderland, MA: Sinauer Associates).

3. Daubin, V., and Moran, N.A. (2004). Comment on" The origins of genome complexity". Science *306*, 978.

4. Kuo, C.H., Moran, N.A., and Ochman, H. (2009). The consequences of genetic drift for bacterial genome complexity. Genome Res. *19*, 1450–1454.

5. Dufresne, A., Garczarek, L., and Partensky, F. (2005). Accelerated evolution associated with genome reduction in a free-living prokaryote. Genome Biol. *6*, R14.

6. Partensky, F., and Garczarek, L. (2010). *Prochlorococcus*: advantages and limits of minimalism. Annu. Rev. Mar. Sci. *2*, 305–331.

7. Batut, B., Knibbe, C., Marais, G., and Daubin, V. (2014). Reductive genome evolution at both ends of the bacterial population size spectrum. Nat. Rev. Microbiol. *12*, 841–850.

8. Marais, G.A., Calteau, A., and Tenaillon, O. (2008). Mutation rate and genome reduction in endosymbiotic and free-living bacteria. Genetica *134*, 205–210.

9. Bourguignon, T., Kinjo, Y., Villa-Martin, P., Coleman, N.V., Tang, Q., Arab, D.A., Wang, Z., Tokuda, G., Hongoh, Y., Ohkuma, M., *et al*. (2020). Increased mutation rate is linked to genome reduction in prokaryotes. Curr. Biol. *30*, XXXX–XXXX.

10. Acosta, S., Carela, M., Garcia-Gonzalez, A., Gines, M., Vicens, L., Cruet, R., and Massey, S.E. (2015). DNA repair is associated with information content in bacteria, archaea, and DNA viruses. J. Hered. *106*, 644–659.

In Brief

Despite intense research on genome architecture since the 2000's, genome-size evolution in prokaryotes has remained puzzling. Using a phylogenetic approach, a new study found that increased mutation rate is associated with gene loss and reduced genome size in prokaryotes.